University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

Gaussian Process Models of Sound Change in Indo-Aryan Dialectology

Cathcart, Chundra

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-184119
Conference or Workshop Item
Published Version

Originally published at:
Cathcart, Chundra (2019). Gaussian Process Models of Sound Change in Indo-Aryan Dialectology. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, Florence, 2019. Association for Computational Linguistics, 254-264.

Gaussian Process Models of Sound Change in Indo-Aryan Dialectology

Chundra A. Cathcart
Department of Comparative Linguistics
University of Zurich
Plattenstrasse 54
CH-8032 Zürich
chundra.cathcart@uzh.ch

Abstract

This paper proposes a Gaussian Process model of sound change targeted toward questions in Indo-Aryan dialectology. Gaussian Processes (GPs) provide a flexible means of expressing covariance between outcomes, and can be extended to a wide variety of probability distributions. We find that GP models fare better in terms of some key posterior predictive checks than models that do not express covariance between sound changes, and outline directions for future work.

1 Introduction and Background

There exists today a wealth of digitized etymological resources from which etymological headwords (e.g., words in Latin, Sanskrit, etc.) and their reflexes in modern language can be extracted, and by proxy, information regarding sound changes operating between ancestral and descendant languages. This information can be used to address hypotheses regarding dialectal relationships between these descendant languages, and the accumulation of large data sets allows such hypotheses to be addressed probabilistically.

This paper builds upon Cathcart to appear, which seeks to address the issue of Indo-Aryan dialect groupings using data extracted from Turner (1962–1966). It has generally been held that sound change holds a great deal of power in determining dialectal relationships in Indo-Aryan (Masica, 1991), and a number of sound changes thought to be probative with respect to Indo-Aryan dialectology have been put forth (Hock, 2016). A problem, however, is that Indo-Aryan languages have developed in close contact with each other, and intimate lexical borrowing between closely related languages has been widespread. Forms showing irregular outcomes of sound change are so great in number that it is difficult to characterize the expected outcomes of sound changes in many languages, much less identify the so-called "residual forms" deviating from what is expected (cf. Bloomfield, 1933).

For this reason, we seek to represent Indo-Aryan languages using a shared-admixture model whereby a given Indo-Aryan language (e.g., Hindi) inherits its vocabulary from multiple LATENT DIALECTAL COMPONENTS in which different SOUND CHANGES have operated; we believe that this approach explicitly models intimate borrowing between Indo-Aryan dialects, a sociolinguistic process that many scholars have argued for (Turner, 1975 [1967]). We restrict the dossier of sound changes we work with to include relatively transparent changes thought to be highly diagnostic for purposes of Indo-Aryan dialectology, with the additional hope of excluding those where multiple intermediate developments have been telescoped into a single change.

The main objective of this paper is to determine the most appropriate way to represent dialect component-level distributions over sound changes. Cathcart to appear compared a shared-admixture model where a Dirichlet prior was placed over sound change probabilities with a model that used a Partitioned Logistic Normal prior, the latter distribution generating Multinomial/Categorical probabilities (like the Dirichlet distribution) but capable of expressing covariance between outcomes within and across distributions (unlike the Dirichlet distribution), and found no major differences in behavior between these two models. At the same time, this procedure relied on a fixed covariance matrix for the Logistic Normal distribution based on the similarity of segments across the sound changes in which they are involved. Working within a similar modeling framework, this paper seeks to model this covariance via a Gaussian Process. Gaussian Processes (GPs) are a flexible family of prior distributions over covari-

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 254–264
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

ance kernel functions. For our purposes, GPs allow us to assess the extent to which sound changes in an evolving linguistic system are correlated, and which features of sound changes influence this correlation. Our results are somewhat open ended at this stage, but we find that GP models fare better in terms of certain critical posterior predictive checks than models that do not express covariance between sound changes.

2 Sound Change

The sound changes that operate within a language's history tend to be subject to certain constraints. In general, most sound changes are thought to stem from low-level phonetic variation, though this view has been challenged (Blust, 2005). Additionally, it is often the case that similar sounds behave similarly in similar environments; hence, if earlier *p* undergoes voicing to *b* between two vowels, it is reasonable expect the changes *t* > *d* and *k* > *g* in the same environment. However, this systematicity and symmetry cannot always be relied upon. Different sounds, regardless of their similarity along a large number of phonetic dimensions, are subject to different articulatory and perceptual constraints. For instance, it is less likely for velar plosives such as *k* to undergo voicing, because considerable articulatory effort is required to pronounce *g* relative to *d* and *b* (Maddieson, 2013). The voiceless labial plosive *p* lacks perceptual salience, and often is debuccalized, losing its oral constriction, to *h* (as in Japanese and Kannada, among other languages) or perceptually enhanced (e.g., to *f*), though other voiceless stops may not undergo the same type of behavior. In other examples, the phonetic grounding is less clear: in most High German dialects, the Old High German consonant shift involved the changes **p* > *(p)f* and **t* > *(t)s*; in southern dialects, the shift also involves the change **k* > *k(x)*; see Schrijver 2014, 97–121 for a sociolinguistic explanation of this asymmetry. In short, while sound change has the tendency to be highly systematic, with similar sounds moving in lockstep, it is clear that this is not always the case; an ideal architecture for modeling sound change will allow for, but not enforce, the possibility of correlation between changes involving similar sounds.

3 Quantitative models of sound change

Under the Neogrammarian view, sound change is a tightly constrained process with discrete binary outcomes; a sound in a given environment has one and only one regular reflex. If irregularity is seen, it is due to analogy or language contact; if neither analogy nor language contact (or, according to some, a small number of additional minor processes that are poorly understood) can be convincingly invoked, then we do not understand the conditioning environment properly. In probabilistic treatments of language change, however, this assumption is infeasible to implement; generally some probability mass, however small, must be allocated to unobserved events (cf. Laplace's law of succession). For this reason, it is standard to relax the Neogrammarian hypothesis by assuming a multinomial/categorical distribution over possible reflexes of a given sound in a language's history (cf. Bouchard-Côté et al., 2007, 2008, 2013); all of the sound changes that operate in the history of a given language can be represented as a collection of multinomial probability distributions, with each distribution in collection corresponding to the possible outcomes of an Old Indo-Aryan (OIA) input in the relevant conditioning environment.

3.1 Prior distributions

In the Bayesian context, an obvious prior for each Multinomial distribution in a collection is the DIRICHLET DISTRIBUTION, which generates probability simplices. The concentration parameter of a SYMMETRIC DIRICHLET DISTRIBUTION can determine the smoothness/sparsity of the resulting multinomial distribution; this is a desirable property, since many phenomena in natural language, sound change being no exception, are best represented using sparse distributions (cf. Ranganath et al., 2015). The Dirichlet distribution has been used to model sound change in previous work (Bouchard-Côté et al., 2007).

However, the Dirichlet lacks an explicit means of expressing correlations between the probabilities of events, such as similar outcomes of sound change, or of modeling dependence between events across multinomial distributions in a collection (like the one we use to represent sound change). An alternative is the LOGISTIC NORMAL DISTRIBUTION (Aitchison, 1986). Under the logistic normal distribution, unbounded values representing unnormalized log probabilities

are generated from a multivariate normal distribution; these are subsequently transformed to probability simplices summing to one via the softmax function. Since the underlying distribution is multivariate normal, the logistic normal distribution is capable of modeling covariance between different outcomes. At the same time, it is not possible to control the sparsity of a logistic normal distribution unless there is high variance and no covariance between different outcomes (this makes it possible to control sparsity in Laplace's approximation to the Dirichlet distribution). Despite this tradeoff, we believe that the logistic normal distribution has promise for modeling sound change, particularly when distributions are noisy. Crucially, the partitioned logistic normal distribution (Cohen and Smith, 2009) allows us to capture dependencies across distributions in a collection as well as within them (i.e., with an eye to modeling low-level variation within dialect groups), allowing us to treat our collection as a large, interdependent distribution.

3.2 Gaussian Processes

Use of the logistic normal distribution in Natural Language Processing usually estimates the covariance between outcomes empirically (cf. Blei and Lafferty, 2007). At the outset, we are unsure of how covariance between two sound changes drawn from a logistic normal prior should be modeled. In principle, covariance should be based on the phonetic similarity of the segments involved, but it is not clear whether all features of all participating segments should have equal influence on the covariance between two changes.

For this reason, we adopt a Gaussian Process approach (Rasmussen and Williams, 2006) to generate our unnormalized sound change probabilities. GPs define a flexible prior over continuous covariance functions. A zero-mean GP assumes that for a given observable response variable, the values of N data points are generated from a multivariate normal distribution with a mean of zero and some covariance. The distribution's covariance is modeled via a kernel function, which takes as its input a measure of distance or dissimilarity between two covarying data points. A popular function is the squared exponential kernel (K_{SE}), which we employ in this paper. A basic squared exponential kernel models the covariance between two data points with values x_i and x_j for some variable in the following manner:

$$K_{\text{SE}}(x_i, x_j) = \alpha^2 \exp\left(\frac{(x_i - x_j)^2}{2\rho^2}\right) \quad (1)$$

The function is parametrized by a parameter α^2, determining the dispersion of the variance-covariance matrix, and a parameter ρ, often referred to as the CHARACTERISTIC LENGTH SCALE, since it controls the distance threshold at which two data points can influence one another, with high values permitting greater influence between distant data points. A third dispersion parameter σ^2 is generally added to diagonal values of the variance-covariance matrix to ensure that it is positive definite. Given a set of data points differing according to a predictor value for which response values are recorded, the parameters α^2, ρ and σ^2 can be fitted conditioned on the data.

We wish to exploit the flexibility of GPs in order to determine how much influence features of segments participating in sound changes should have on other coextensive sound changes. Take the changes $p > b$ and $t > d$, setting aside the conditioning environment. Both straightforwardly involve voicing of a voiceless plosive. Care must be taken in representing these changes in a way that the relevant dimensions of similarity can be detected by a probabilistic model. If we compute similarity between them on the basis of whether the segments involved are identical, we will not be able to take into account processes such as voicing — i.e., $p > b$ and $t > d$ (which both involve voicing) will be treated as being as dissimilar as $p > b$ and $d > t$ (which involve voicing and devoicing, respectively). Such a model may not be completely useless, as it will still capture correlations between identical changes across different environments, a generalization that the Dirichlet distribution is not explicitly capable of capturing.

In contrast to a binary approach concerned with segmental identity, we can make use of distinctive phonological features to capture granular relationships between similar sound changes. If we assume a simple featural representation for each change, these changes will differ along the dimension of PLACE OF ARTICULATION (`labial` > `labial` ≠ `dental` > `dental`) but not VOICING (`voiceless` > `voiced` = `voiceless` > `voiced`).

We are faced with similar questions when deciding how to represent the conditioning environ-

ment. While it makes sense that the featural representations of the input and output of each individual change should be considered jointly, it is not clear that the environment should be treated in such a manner. If we look only at the joint dissimilarity of the lefthand and righthand contexts of each pair of changes, there is the potential that the dissimilarity between changes where only one side of the environment is a relevant conditioning factor will be inflated if the other side differs. Therefore it may be more instructive to model similarity between conditioning environments as a composition of the similarities of the left- and righthand contexts, though this model may have the potential to overgeneralize. We opt to treat the environment as a whole as a feature of interest, based on a survey of conditioning environments (Kümmel, 2007), setting this question aside for future work.

There are several ways to deal with multiple variables or featural dimensions in a GP framework. The simplest approach is to assume a single length scale for all features, which can potentially induce behavior similar to an interaction in a linear model — if the length scale is low, covariance between two data points will be high only if their similarity across all dimensions is high as well. An alternative is to assume a kernel function for each dimension $d \in \{1, ..., D\}$, and add these together. A third approach is to model an additive combination of the dimensions within the kernel function, as follows:

$$K_{\text{SE}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \alpha^2 \exp\left(\sum_{d=1}^{D} \frac{(x_{i,d} - x_{j,d})^2}{2\rho_d^2}\right) \quad (2)$$

A consequence of the structure of this kernel, known as an Automatic Relevance Determination (ARD) kernel, is that covariance will not be sensitive to or vary according to differences along dimensions for which ρ_d is large, allowing us to gauge which featural dimensions have greater "relevance" (Neal, 1996). While interpreting relevance is challenging for featural dimensions which have different scales (Piironen and Vehtari, 2016), this is not a concern for our data, since distances between sound changes across featural dimensions are binary (i.e., 0 or 1).

We employ an ARD kernel for two types of GP prior over covariance between sound changes. The first kernel, the binary GP (BGP) takes into account two dimensions concerning (1) segmental identity between inputs and outputs and (2) segmental identity between environments across each pair of changes. The granular GP (GGP) generalizes this approach to a larger number of dimensions corresponding to phonological features of interest, described below.

3.3 Feature representation and kernel structure

We assume an n-ary featural representation for the sound types in our data set, similar to that found in models such as that of Futrell et al. (2017). In theory, it would be possible to employ binary distinctive features à la *The Sound Pattern of English* (Chomsky and Halle, 1968) and related works, which would potentially allow a richer representation (Duvenaud, 2014), but with considerable computational cost. Embedding representations for continuous phonetic values present a promising avenue (cf. Cotterell and Eisner, 2017). The feature space looks as follows:

- A feature indicating whether a segment is a CONSONANT or VOWEL
- A set of consonant-specific features:
  - Place of articulation: labial, dental, palatal, retroflex, velar, glottal
  - Manner of articulation: plosive, affricate, fricative, approximant, nasal
  - Voicing: ±
  - Aspiration: ±
- A set of vowel-specific features:
  - Height: low, mid, high
  - Frontness: front, back
  - Rounding: ±
  - Orality: oral, nasal

This yields 9 featural dimensions. Each segment takes an n-ary or binary value for each relevant attribute; for irrelevant attributes (i.e., consonant-specific features, if the segment is a vowel, or vice versa), the segment is assigned a null value.

4 Data

We extracted all modern Indo-Aryan (NIA) forms from Turner (1962–1966) along with the OIA headwords from which these reflexes descend (Middle Indo-Aryan languages such as Prakrit and Pali were excluded). Transcriptions of the data

were normalized and converted to the International Phonetic Alphabet (IPA). Systematic morphological mismatches between OIA etyma and reflexes were accounted for, including stripping the endings from all verbs, since citation forms for OIA verbs are in the 3sg present, while most NIA reflexes give the infinitive. We matched each dialect with corresponding languoids in Glottolog (Hammarström et al., 2017) containing geographic metadata, resulting in the merger of several dialects. Languages with fewer than 100 forms in the data set were excluded, yielding 50 remaining languages; the best represented language is Hindi, with 4012 forms, followed by Sinhala, Marathi, Panjabi and Gujarati. We excluded sound changes appearing fewer than 7 times in our data set, ultimately yielding 38479 modern Indo-Aryan words. We preprocessed the data, first converting each segment into its respective sound class, as described by List (2012), and subsequently aligning each converted OIA/NIA string pair via the Needleman-Wunsch algorithm, using the Expectation-Maximization method described by Jäger (2014), building off of work by Wieling et al. (2012). This yields alignments of the following type: e.g., OIA /aːntra/ 'entrails' > Nepali /aːnØro/, where Ø indicates a gap where the "cursor" advances for the OIA string but not the Nepali string. Gaps on the OIA side are ignored, yielding a one-to-many OIA-to-NIA alignment; this ensures that all aligned cognate sets are of the same length. We restrict our analysis to changes affecting OIA ʃ, ʊ, ɲ, ɳ, ʂ, ɾ, h, i, iː, j, kʂ, l, n, r, s, u, uː, which are thought to play a meaningful role in Indo-Aryan dialectology (Southworth, 2005; Hock, 2016).

5 Model

Complete information regarding this paper's model specification and inference can be found in the Appendix. Our data set contains W OIA etyma, each of which is continued by some of the L languages in our sample. The data set contains R OIA inputs (e.g., sounds in a conditioning environment), each of which have S_r reflexes. We assume $K = 10$ dialect groups. At a high-level, our model is a mixed membership model which assumes that EACH WORD in EACH LANGUAGE is generated by one of K latent dialect components, according to the relevant sound changes whose operation the word displays. Key parameters are $\boldsymbol{\theta}$ (language-level distributions over dialect components) and $\boldsymbol{\phi}$ (component-level collections of distributions over sound changes). The stochastic generative process we assume to underlie the data looks as follows (for information regarding priors over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, refer to the Appendix):

For $\boldsymbol{w}_i : i \in \{1, ..., W\}$, the vector of relevant inputs in each OIA etymon

  For each language $l \in \{1, ..., L\}$ continuing $\boldsymbol{w}_i$

    $z_{i,l} \sim$ Categorical$(\boldsymbol{\theta}_l)$ [Draw a dialect component label]

    For each OIA input $w_{i,t}$ in etymon $\boldsymbol{w}_i$ at index $t : \{1, ..., |\boldsymbol{w}_i|\}$

      $y_{i,l,t} \sim$ Categorical$(\boldsymbol{\phi}_{z_{w,l}, w_{i,t}, \cdot})$ [Generate each output]

The likelihood of a given NIA word's reflexes (i.e., outcomes of relevant sound changes) $\boldsymbol{y}_{i,l}$ and its OIA predecessor $\boldsymbol{w}_i$ under the generative process described above is the following, with the discrete variable $z_{i,l}$ marginalized out:

$$P(\boldsymbol{y}_{i,l}, \boldsymbol{w}_i | \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \theta_{l,k} \prod_{t=1}^{|\boldsymbol{w}_i|} \phi_{k, w_{i,t}, y_{i,l,t}} \quad (3)$$

We carry out inference for three flavors of this model involving different versions of $\boldsymbol{\phi}$. In the Diagonal model, there is no covariance across outcomes of $\boldsymbol{\phi}$. In the Binary GP (BGP) and Granular GP (GGP) models, $\boldsymbol{\phi}$ is generated by GPs with the ARD kernels described in 3.2; these models differ in that the former takes a 2-dimensional featural input, while the latter takes a 18-dimensional one (2 times the number of features given in 3.3). We fit a variational posterior to the data for multiple separate initializations (as described in the Appendix) from which we can draw samples.

6 Results

6.1 Geographic distribution

Averaged language-level component distributions can be visualized geographically in Figure 1. A number of redundant components are shared across all languages in each model; this is likely an artifact of the prior placed over $\boldsymbol{\theta}$; changes to this prior (see discussion in the Appendix) would likely assign less probability mass to redundant components. In general, for all models, certain linguistic groups show a similar component makeup: these groups include Romani dialects and their

close relatives Domari and Lomavren; Dardic languages of northern Pakistan; languages of Eastern South Asia and the Eastern Indo-Gangetic Plain; the insular languages Sinhala and Dhivehi; and western languages such as Marathi and Gujarati.

We measure correlation coefficients to assess how well the language-level dialect component makeup inferred in each of our models reflects the geography of Indo-Aryan dialects. For each of our three models, we compute the Jensen-Shannon divergence between $\boldsymbol{\theta}_l$ and $\boldsymbol{\theta}_{l'}$ for each pair of languages l, l', averaging across samples of $\hat{\boldsymbol{\theta}}$, the language-level posterior over components. We measure the correlation between (1) average inter-language JS divergence between dialect component makeup and (2) pairwise great circle geographic distance, using Spearman's ρ (although pairwise distances violate the independence assumption). These values are .28 for the Diagonal model, .34 for the BGP model, and .26 for the GGP model. We see that the BGP model shows the strongest geographic signal. We note that this metric serves as a basis for comparison, but not evaluation; if the language contact we are detecting is chronologically deep, it is less likely to show a strong geographic signal (cf. Haynie, 2012).

6.2 Relevance

We inspect posterior values of $\boldsymbol{\rho}^{-2}$, the squared inverse characteristic length scales for each featural dimension of interest, for both the BGP and GGP models. Since we work with inverse scales, high values indicate relevance, while values close to zero indicate irrelevance.

Figure 2 shows the squared inverse length scales for the BGP model. The squared inverse length scale for change is higher than that of environment, though the multimodality seen may be due to a lack of convergence across initializations for the BGP model. This is perhaps not particularly surprising, though perhaps something of a sanity check: given a large number of sound changes involving a large number of conditioning environments, some of them redundant, it is likely that changes with different environments and identical input-output pairs will show similar behavior.

Figure 3 shows the squared inverse length scales for the GGP model. The results seem to suggest that when input-output pairs and conditioning environments are decomposed into featural representations, very few featural dimensions have a strong influence on the co-occurrence of sound changes that show featural identity in terms of input-output pair or conditioning environment — essentially, these features are the most meaningful when they are bundled together into individual segments. An exception is the feature VOWEL HEIGHT for environment, indicating that changes are likely to co-occur if their conditioning environments have the same values for vowel height. Further work is needed to determine which combination of feature values for the left- and righthand context in the conditioning environments actually serves as a meaningful determinant of correlation.

6.3 Posterior Predictive Checks

6.3.1 Entropy

We carry out model criticism using a posterior predictive check proposed by Mimno et al. (2015) for mixed-membership models, inspecting the uncertainty with which each model assigns dialect component labels to each word. Recall that during inference, we marginalized out the discrete variables $z_{i,l}$, which indicate the dialect component label selected for the reflex of OIA word w_i in language l. Given our fitted parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$, it is straightforward to reconstruct the probability of a label for a given NIA word:

$$P(z_{i,l} = k|\boldsymbol{y}_{i,l}, \boldsymbol{w}_i, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \theta_{l,k} \prod_{t=1}^{|\boldsymbol{w}_i|} \phi_{k,w_{i,t},y_{i,l,t}} \tag{4}$$

If $P(z_{i,l}|\boldsymbol{y}_{i,l}, \boldsymbol{w}_i, \boldsymbol{\theta}, \boldsymbol{\phi})$ shows high entropy, then our fitted parameters do not allow us to assign a label with certainty. We average the entropy of $P(z_{i,l}|\boldsymbol{y}_{i,l}, \boldsymbol{w}_i, \boldsymbol{\theta}, \boldsymbol{\phi})$ across each word for 100 samples of $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$ in each model. Histograms of these entropy measures can be seen in Figure 4. The averages of these averaged values are 1.058 for the Diagonal model, 1.255 for the BGP model, and 1.259 for the GGP model, with the Diagonal model outperforming the GP models; these values show a decrease rather than an increase in posterior predictive checks with greater granularity in the underlying GP.

6.3.2 Accuracy

We assess the extent to which each model's posterior parameters can accurately regenerate the observed data. For each word, we sample $z_{i,l} \sim \text{Categorical}(\hat{\boldsymbol{\theta}}_l)$, and then draw outcomes of sound change $\hat{y}_{i,l,t} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{i,l},w_{i,t},\cdot}) : t \in$

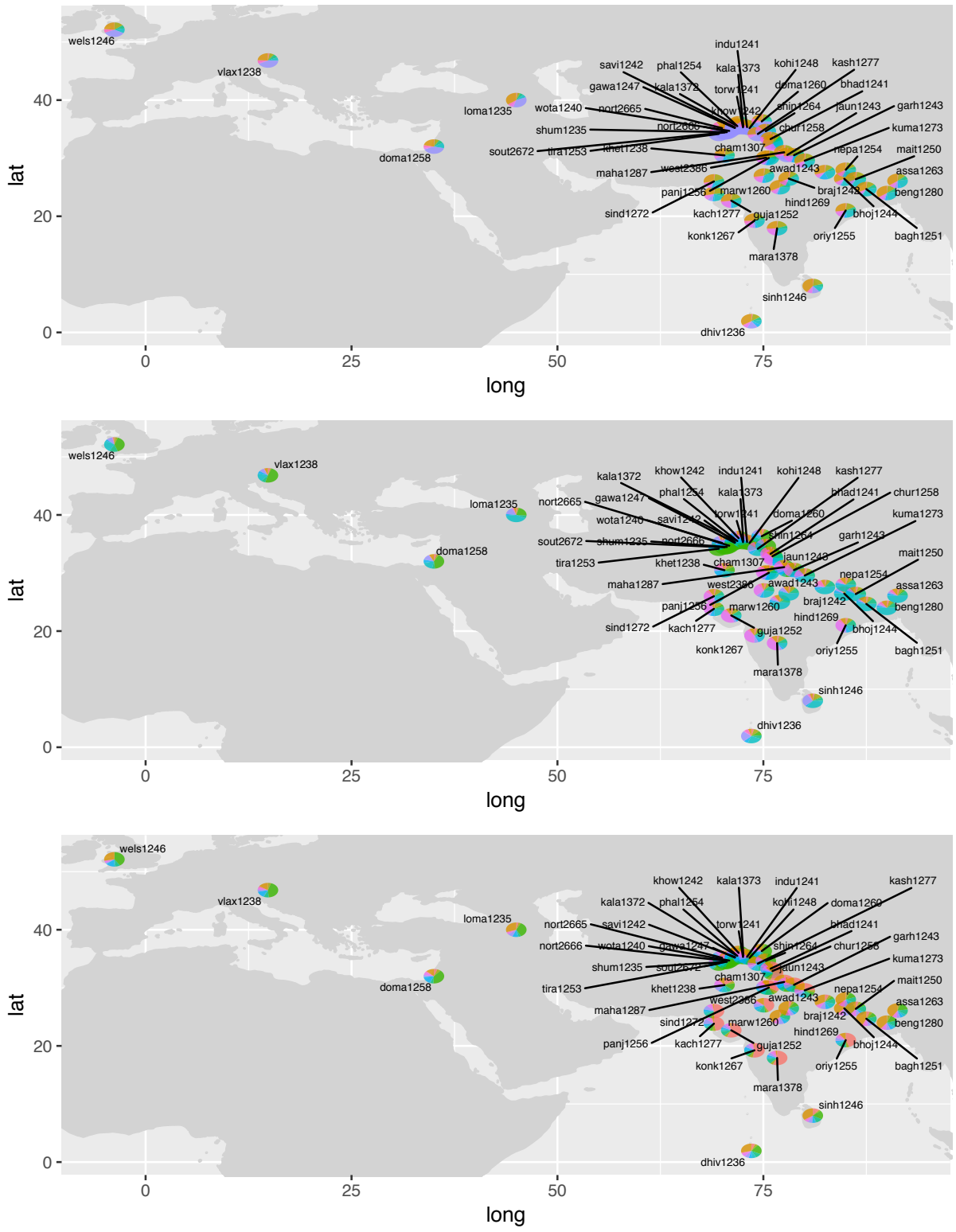

Figure 1: Averaged language-level component distributions for Diagonal (top), BGP (middle), and GGP (bottom) models.

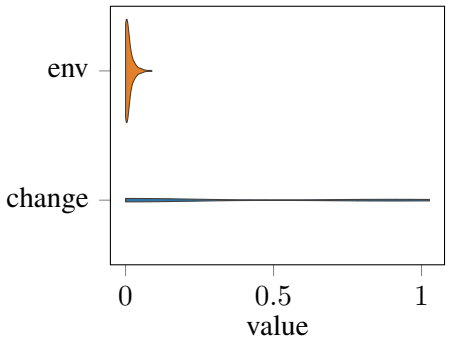

Figure 2: Squared inverse scales for the BGP model by featural dimension.

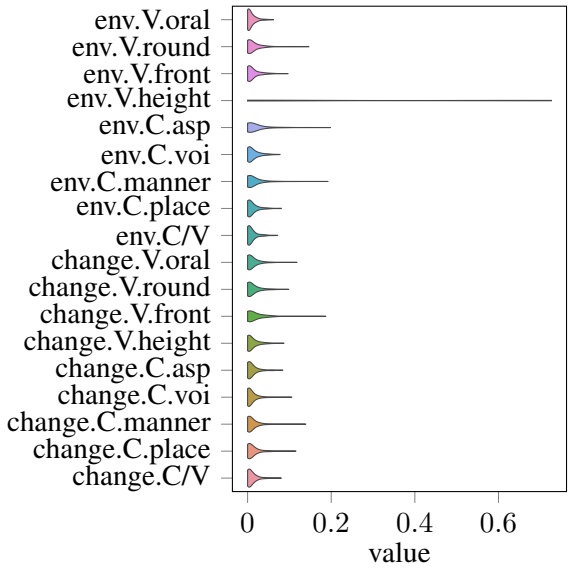

Figure 3: Squared inverse scales for the GGP model by featural dimension.

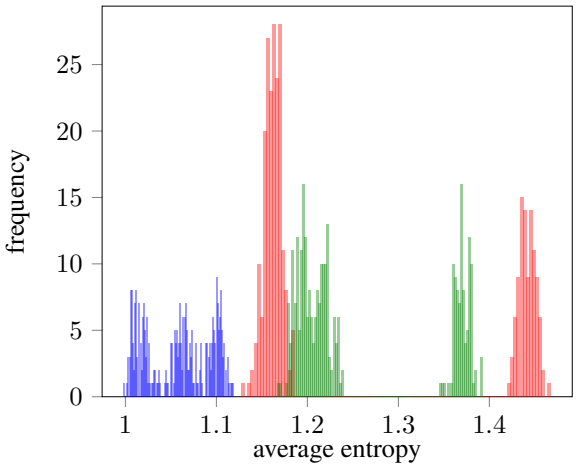

Figure 4: Average word-level component assignment entropies from posterior samples for each model (Diagonal = blue, BGP = red, GGP = green).

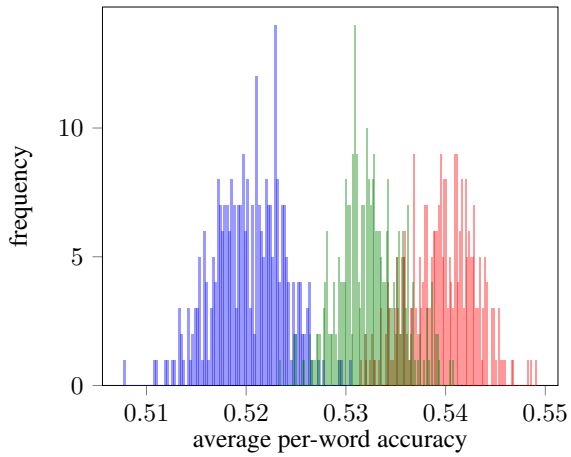

Figure 5: Average per-word accuracies from posterior samples for each model (Diagonal = blue, BGP = red, GGP = green).

$\{1, ..., |\boldsymbol{w}_i|\}$. We measure per-word accuracy by dividing the number of instances of $\hat{y}$ that were correctly simulated by the number of relevant sound changes in the word, $|\boldsymbol{w}_i|$. We take the mean of per-word averages across the data set for 100 samples of $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$ in each model. Histograms of these accuracy measures can be seen in Figure 5. We find that the GP models re-generate the data with greater accuracy than the Diagonal model, but the BGP model outperforms the GGP model. This suggests that the sound change posterior distributions $\hat{\boldsymbol{\phi}}$ of the GP models are more informative than those of the Diagonal model, and better capture the structure of the data. It is possible that the Diagonal model fared better in terms of entropy due to a trade-off in sparsity between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$, where more informative $\hat{\boldsymbol{\theta}}$ and flatter $\hat{\boldsymbol{\phi}}$ allowed for component labels to be assigned with greater certainty.

7 Outlook

This paper proposed a probabilistic formalization of sound change according to the logistic normal distribution, a distribution that has been underused for such a modeling purpose. We attempted to use GPs in order to induce more realistic sound change distributions for application to dialectological questions. We described a generative Bayesian model in which unnormalized logistic normal weights are generated by a Gaussian Process, a powerful and flexible prior distribution over functions that can be used to model covariance for multivariate normal data. GPs have been put forth as a means of modeling continuous pho-

netic changes (Aston et al., 2012), but this paper is the first to use them as a prior for multinomial sound change distributions.

While some aspects of our results were difficult to interpret and remain inconclusive, we did demonstrate a marginal increase in terms of key posterior predictive checks with the use of Gaussian Process models. It is clear that much work is required in order to bring the automated methodology described here into line with gold standards in linguistics as well as the intuitions of historical linguistics; however, we believe that this research program is promising and has high potential impact. Specifically, received wisdom can be used in the process of prior selection for Bayesian models. In this paper, we used a standard and simple covariance kernel function for our Gaussian process, the squared exponential kernel. We placed relatively uninformative priors over the parameters of the kernel function in the hopes that well-informed, highly identifiable parameters would fall out of the data. Further empirical work is required to determine which priors over kernel parameters are suitable, if a squared-exponential kernel is to be used in future work. Additionally, it is worth noting that there are many kernel functions to choose from, and that the squared-exponential kernel has its limitations. It (along with many other popular functions used for GPs) cannot model negative covariance, for example, whereas highly sophisticated alternatives can (Wilson and Adams, 2013).

If the methodology described here can be refined, the potential for quantitative historical linguistics is significant. Sound change and morphological change are the cornerstones of traditional historical linguistics (Meillet, 1922). High-definition data sets like the one used in this paper are largely unexploited. If the issues outlined above can be tackled, models like the one employed in this paper will undoubtedly serve as a powerful means of inferring key aspects of linguistic prehistory.

References

John Aitchison. 1986. *The statistical analysis of compositional data*. Chapman & Hall, London/New York.

JAD Aston, D Buck, J Coleman, CJ Cotter, NS Jones, V Macaulay, N MacLeod, JM Moriarty, and A Nevins. 2012. Phylogenetic inference for function-valued traits: speech sound evolution. *Trends in Ecology & Evolution*, 27(3):160–166.

David M Blei and John D Lafferty. 2007. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.

Robert A Blust. 2005. Must sound change be linguistically motivated? *Diachronica*, 22(2):219–269.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110:4224–4229.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague. Association for Computational Linguistics.

Alexandre Bouchard-Côté, Percy S Liang, Dan Klein, and Thomas L Griffiths. 2008. A probabilistic approach to language change. In *Advances in Neural Information Processing Systems*, pages 169–176.

Chundra Cathcart. to appear. A probabilistic assessment of the Indo-Aryan Inner-Outer Hypothesis. *Journal of Historical Linguistics*.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Haper & Row, Publishers, New York.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82. Association for Computational Linguistics.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192.

David Duvenaud. 2014. *Automatic Model Construction with Gaussian Processes*. Ph.D. thesis, University of Cambridge.

Richard Futrell, Adam Albright, Peter Graff, and Timothy J. O'Donnell. 2017. A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, 5:73–86.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.3. Max Planck Institute for the Science of Human History.

Hannah Haynie. 2012. *Studies in the History and Geography of California Languages*. Ph.D. thesis, University of California, Berkeley.

Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. 2012. Kernel topic models. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *JMLR*, La Palma, Canary Islands.

Hans Henrich Hock. 2016. The languages, their histories, and their genetic classification. In Hans Henrich Hock and Elena Bashir, editors, *The Languages and Linguistics of South Asia: A Comprehensive Guide*, pages 9–240. De Gruyter, Berlin, Boston.

Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155–204. Brill.

Mohammad Khan, Shakir Mohamed, Benjamin Marlin, and Kevin Murphy. 2012. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *Artificial Intelligence and Statistics*, pages 610–618.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Adam Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Martin Kümmel. 2007. *Konsonantenwandel*. Dr. Ludwig Reichert Verlag, Wiesbaden.

Johann-Mattis List. 2012. SCA. Phonetic alignment based on sound classes. In M. Slavkovik and D. Lassiter, editors, *New directions in logic, language, and computation*, pages 32–51. Springer, Berlin, Heidelberg.

Ian Maddieson. 2013. Voicing and gaps in plosive systems. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Colin P. Masica. 1991. *The Indo-Aryan languages*. Cambridge University Press, Cambridge.

Antoine Meillet. 1922. *Les dialectes indo-européens*. E. Champion, Paris.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419. JMLR. org.

David Mimno, David M. Blei, and Barbara E. Engelhardt. 2015. Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences*, page 201412301.

Radford Neal. 1996. *Bayesian Learning for Neural Networks*. Springer, Berlin and Heidelberg.

J. Piironen and A. Vehtari. 2016. Projection predictive model selection for Gaussian processes. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE*.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. 2015. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771.

C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and variational inference in deep latent gaussian models. *arXiv preprint arXiv:1401.4082*.

Peter Schrijver. 2014. *Language contact and the origin of Germanic languages*. Routledge, New York.

Franklin C. Southworth. 2005. *Linguistic Archaeology of South Asia*. Routledge, London.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations (ICLR)*.

Ralph L. Turner. 1962–1966. *A comparative dictionary of Indo-Aryan languages*. Oxford University Press, London.

Ralph L. Turner. 1975 [1967]. Geminates after long vowel in Indo-aryan. In *R.L. Turner: Collected Papers 1912–1973*, pages 405–415. Oxford University Press, London.

Martijn Wieling, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.

Andrew Wilson and Ryan Adams. 2013. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta. https://arxiv.org/pdf/1302.4245.pdf.

8 Appendix

Here, we describe our model specification as well as the inference procedure used to fit our model's parameters.[1] We parameterize our model such that

[1] All relevant code can be found at https://github.com/chundrac/IA_dial/tree/master/LChange2019.

no random variables are dependent on other random variables, treating such variables as deterministic variables dependent on an auxiliary noise variable and one or more random variables. This allows us to construct a straightforward variational approximation to our model.

8.1 Language-component prior

The parameter $\boldsymbol{\theta}$, representing language-level distributions over latent dialect components, is generated as follows:

$$\eta_{l,k} \sim \mathcal{N}(0, 10) : l \in \{1, ..., L\}, k \in \{1, ..., K\}$$

$$\boldsymbol{\theta}_l = \text{softmax}(\boldsymbol{\eta}_l)$$

Placing a large standard deviation on the Gaussian prior passed to the softmax function allows for sparser multinomial distributions to be generated, but unlike symmetric Dirichlet priors with a concentration parameter below 1, does not penalize smoother distributions relative to sparse ones.

In theory, the Gaussian Stick-Breaking construction of (Khan et al., 2012; Miao et al., 2017) can be used to allow the language-level prior over dialect components to favor a large or small number of groups, conditional on the data. We do not use the GSB prior in this paper, but are exploring it in ongoing work.

8.2 Component-sound change prior

8.2.1 Diagonal Prior

The diagonal prior (i.e., the prior over sound changes that is insensitive to correlation) is a softmax-transformed diagonal multivariate Gaussian distribution with high variance:

$$\psi_{k,r,s} \sim \mathcal{N}(0, 10) : k \in \{1, ..., K\}, \\ r \in \{1, ..., R\}, s \in \{1, ..., S_r\}$$

$$\boldsymbol{\phi}_{k,r,\cdot} = \text{softmax}(\boldsymbol{\psi}_{k,r,\cdot})$$

8.2.2 GP Prior

The following process holds for both the Binary GP (BGP) and Granular GP (GGP), the only difference being that the dimensionality D of the $D \times S \times S$ matrix δ containing pairwise featural distances between sound changes is larger for the GGP model. We use the Cholesky decomposition of the variance-covariance matrix $\boldsymbol{\Sigma}$ generated by the SEK function, coupled with an auxiliary noise variable, in order to treat $\boldsymbol{\psi}$ as a deterministic random variable.

$$\alpha \sim \mathcal{N}(0, 1), \sigma \sim \mathcal{N}(0, 10) \\ \rho_d^{-1} \sim \mathcal{N}(0, .1) : d \in \{1, ..., D\}$$

[2]

$$\boldsymbol{\Sigma} = \alpha^2 \exp\left(-\sum_{d=1}^{D} \frac{\delta^d}{2\rho^2}\right) + \boldsymbol{I}\sigma^2 = \boldsymbol{L}\boldsymbol{L}^\top$$

$$z^{\Sigma}_{k,r,s} \sim \mathcal{N}(0, 1) : k \in \{1, ..., K\}, \\ r \in \{1, ..., R\}, s \in \{1, ..., S_r\}$$

$$\boldsymbol{\psi}_{k,\cdot,\cdot} = \boldsymbol{L}\boldsymbol{z}^{\Sigma}_k$$

$$\boldsymbol{\phi}_{k,r,\cdot} = \text{softmax}(\boldsymbol{\psi}_{k,r,\cdot})$$

8.3 Inference

We use Stochastic Gradient Variational Bayes (Kingma and Welling, 2014) to learn each model's parameters. Since all of our priors are Gaussian, it is straightforward to construct a Gaussian variational approximation for each parameter with its own trainable mean and standard deviation. The objective of Variational Inference is to maximize the evidence lower bound (ELBO), given below:

$$\text{ELBO} = \mathbb{E}_{z \sim q(z|x)}\left[P(x|z)\right] - D_{KL}(q(z|x)||p(z))$$

where the first term denotes the expectation of the model log likelihood (see eq. 3) under samples z from the variational posterior $q(z|x)$, and the second denotes the sum of Kullback-Leibler (KL) divergences between the variational posterior parameters and their corresponding priors in $p(z)$, all of which are Gaussian. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of .1 to optimize the variational parameters for 5000 iterations over 3 separate initializations via batch inference (i.e., fitting the parameters on the entire dataset at each iteration), using 10 Monte Carlo samples per iteration to estimate $\mathbb{E}_{z \sim q(z|x)}\left[P(x|z)\right]$ according to the reparameterization trick (Rezende et al., 2014; Kingma and Welling, 2014). To deal with label switching across initializations, we choose the permutation of labels {1,...,K} of the posterior parameters of initializations 2 and 3 such that the KL divergence to the posterior parameters of the first initialization is minimized.

[2] In theory, a more informative prior over σ such as $\mathcal{N}(10, 1)$ may be a good choice in order to encourage sparser distributions.